信息论笔记

宋佳欢

2019 年 9 月 24 日

目录

1 信息熵

信息量 $I(x)=f\big(p(x)\big)$，函数 f 需满足下列四个条件:

1.f 单调递减，事件发生的概率越小，获得的信息量越大。

2. 当 $p(x)=1$，$f\big(p(x)\big)=0$

3. 当 $p(x)=0$，$f\big(p(x)\big)=\infty$

4. 两件独立事件同时发生的获取的信息之和为 $I(x,y) = I(x)+I(y) = f\big(p(x)\big)+f\big(p(y)\big)=f\big(p(x,y)\big)$

因此，$p(x,y)=p(x)p(y)$。根据这个关系，$I(x)$ 与 $p(x)$ 一定为对数关系。

根据上述四个条件可得:

$$I(x)=-logp(x)$$

其中负号是用来保证信息量是正数或者零。而 log 函数基的选择是任意的（信息论中基常常选择为 2，因此信息的单位为比特 bits，即信息需要的编码长度；而机器学习中基常常选择为自然常数，因此单位常常被称为奈特 nats；底数为 10，单位则为 Hart）。

$I(x)$ 也被称为随机变量 x 的自信息 (self-information)，描述的是随机变量的某个事件发生所带来的信息量。

现在假设一个发送者想传送一个随机变量的值给接收者。那么在这个过程中，他们传输的平均信息量可以通过求 $I(x)$ 关于概率分布 $p(x)$ 的期望求得，随机变量 X 的信息熵的定义：

$$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$$

熵越大，随机变量的不确定性就越大。是对所有可能发生的事件产生的信息量的期望。

1.1 代数性质

1.1.1 对称性

变量 $p_1, p_2, \cdots, p_r$ 的顺序任意互换，熵不变。

$$H(p_1, p_2, \cdots, p_r) = H(p_2, \cdots, p_r, p_1) = H(p_r, p_1, p_2, \cdots, p_{r-1})$$

1.1.2 非负性

$$H(p_1, p_2, \cdots, p_r) \geq 0$$

1.1.3 连续性

$H(p_1, p_2, \cdots, p_r)$ 是 p_i 的连续函数。

1.1.4 扩展性

$$\lim_{\varepsilon \to 0} H(p_1, p_2, \cdots, p_i - \varepsilon, \cdots, p_r, \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k) = H(p_1, p_2, \cdots, p_r)$$

其中每个 ε 都趋于 0。当信源的消息集中的消息数增多时，因为这些消息对于的概率很小 (比重很小)，所以信源的熵不变。

1.1.5 可加性

统计独立的两个信源 X,Y 的两个联合信源的熵等于分别熵之和:

$$H(X, Y) = H(X) + H(Y)$$

$$\begin{aligned}
H(X, Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j log p_i q_j \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j log p_i - \sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j log q_j \\
&= -\sum_{i=1}^{n} \big(\sum_{j=1}^{m} p_i q_j\big) log p_i - \sum_{j=1}^{m} \big(\sum_{i=1}^{n} p_i q_j\big) log q_j \\
&= -\sum_{i=1}^{n} p_i log p_i - \sum_{j=1}^{m} q_j log q_j = H(X) + H(Y)
\end{aligned}$$

1.1.6 递增性

将信源 X 中的其中一个元素划分成 m 个元素，这 m 个元素的概率之和等于原元素的概率，熵增加。

$$H_{n+m-1}(p_1,p_2,\cdots,p_{n-1},q_1,q_2,\cdots,q_m)=H_n((p_1,p_2,\cdots,p_{n-1},p_n)+p_nH_m(\frac{q_1}{p_n},\frac{q_2}{p_n},\cdots,\frac{q_m}{p_n})$$

1.2 解析性质

1.2.1 最大离散熵定理

在离散信源情况下, 信源各符号等概率分布时, 熵达到最大。(概率分布越接近平均分布, 熵越大)

$$H(p_1,p_2,\cdots,p_r)\leq H(\frac{1}{r},\frac{1}{r},\cdots,\frac{1}{r})\leq logr$$

1.2.2 上凸性

熵函数是严格的上凸函数:

$$H(\theta P_1+(1-\theta)P_2)>\theta H(P_1)+(1-\theta)H(P_2)$$

2 信道

离散单符号信道可用传递概率表示:

$$\begin{bmatrix} P(b_1|a_1) & P(b_2|a_1) & \cdots & P(b_s|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & \cdots & P(b_s|a_2) \\ \vdots & \vdots & & \vdots \\ (b_1|a_r) & P(b_2|a_r) & \cdots & P(b_s|a_r) \end{bmatrix}$$

a 为输入，b 为输出。且传递矩阵 (信道矩阵) 每一行的元素相加等于 1，即:

$$\sum_{j=1}^{s}P(b_j|a_i)=1$$

$P(b_i|a_i)$ 表示发送 a 收到 b 的概率（前向概率）描述了信道噪声的特征，$P(a_i|b_i)$ 表示接受到了 b_i，发送端发送 a_i 的概率（后向概率）。

2.1 互信息与平均互信息

互信息 (Mutual Information):$I(a_i;b_j)$ 表示接受到 b_j 后，能从 b_j 获得的关于 a_i 的信息量。互信息的三种写法:

$$\begin{aligned} I(a_i;b_j) &= I(a_i)-I(a_i|b_j) \\ &= I(b_j)-I(b_j|a_i) \\ &= I(a_i)+I(b_j)-I(a_i,b_j) \end{aligned}$$

但是单个样本的互信息不足以表示整个系统，因此需要对多个样本取期望，即平均互信息：

$$I(X;Y)=E_{P(X,Y)}\{I(a_i;b_j)\}$$

对于单个样本的互信息，其值可正可负或为零，但是平均互信息一定不会为负值。

证明：

$$I(X;Y)=E_{P(X,Y)}\{I(a_i;b_j)\}=\sum_i\sum_j P(a_i,b_j)log\frac{P(a_i,b_j)}{P(a_i)P(b_j)}$$

$$-I(X;Y)=\sum_i\sum_j P(a_i,b_j)log\frac{P(a_i)P(b_j)}{P(a_i,b_j)}$$

根据琴生不等式，以及：

$$\sum_i\sum_j P(a_i)P(b_j)=1$$

所以

$$-I(X;Y)=\sum_i\sum_j P(a_i,b_j)log\frac{P(a_i)P(b_j)}{P(a_i,b_j)}\leq log\Big(\sum_i\sum_j P(a_i,b_j)\frac{P(a_i)P(b_j)}{P(a_i,b_j)}\Big)=log1=0$$

即

$$I(X;Y)\geq 0$$

互信息有三种写法，平均信息也衍生出三种写法：

$$\begin{aligned}I(X;Y)&=H(X)-H(X|Y)\\&=H(Y)-H(Y|X)\\&=H(X)+H(Y)-H(XY)\end{aligned}$$

其中 $H(X|Y)$ 称为疑义度,$H(Y|X)$ 称为噪声熵，$H(XY)$ 称为联合熵（共熵）。